Machine Learning
Programming Assignment 5

Vaibhav Jindal
111701029

- Learned the theory behing Density Based Clustering
- Learned the restriction and uses of DBSCAN
  - RESTRICTION
    - No need to find k (number of clusters)
    - Can find non-sperical and non-elipsoid and intersecting clusters
    - Finds outliers (abnormal points), not sensitive to outliers like k-means and spectal clustering

  - Uses
    - Problem with selection of hyperparameters i.e. epsilon and min_points
    - Problem with data with varying density
    - Problem in high dimensional data

- Got few methods of finding hyperparameters from input dataset
- Implementing DBSCAN using sklearn

PROBLEM: This assignment is to find out anomalous samples,The definition of anomalous is something abnormal or deviating from the usual groups, Normal samples are in also various groups but the samples which are not belongs to any groups are considered to be abnormal.

OBSERVATIONS

- For finding good epsilon (Radius)
  - As the data is big dimensional in terms of features there must be a way to find epsilon relatable to Radius
  - We have found pairwise distance of each point with the other in terms of Frobenius norm
  - Now epsilon = average of all the pairwise distance
  - This method seems to be fine because points belonging to same cluster should have all points around the average distance of it’s core point

- For finding good min_no_points
  - Within the radius there should be some min number of points to make it a core point,
  - **Observed that input vector have less variance between it's values, Ex input[0] = [65, 67, 67, 61, 55, 57, 66, 74, 64, 69, 61, 62, 71, 71, 53, 53, 64, 71,**
  - **65, 66, 71, 73, 65, 70, 47, 39, 80, 75, 67, 64, 72, 66, 53, 55, 63, 64,**
  - **60, 61, 74, 63, 59, 58, 50, 49]**
    - **Values are between 40-80. Not much variance.**

  - Now we have find the Frobenius norm of each input vector (data point), Points with near Frobenius value should be in one cluster (Given that there is not much variance).
  - Hence I have found the frequency of number of clusters with near norm values.
  - Putting min_no_points = 4*freq

Please refer to code for more clearity.

Time complexity for finding

1) Epsilon = O(n^2)
2) min_no_samples = O(n)

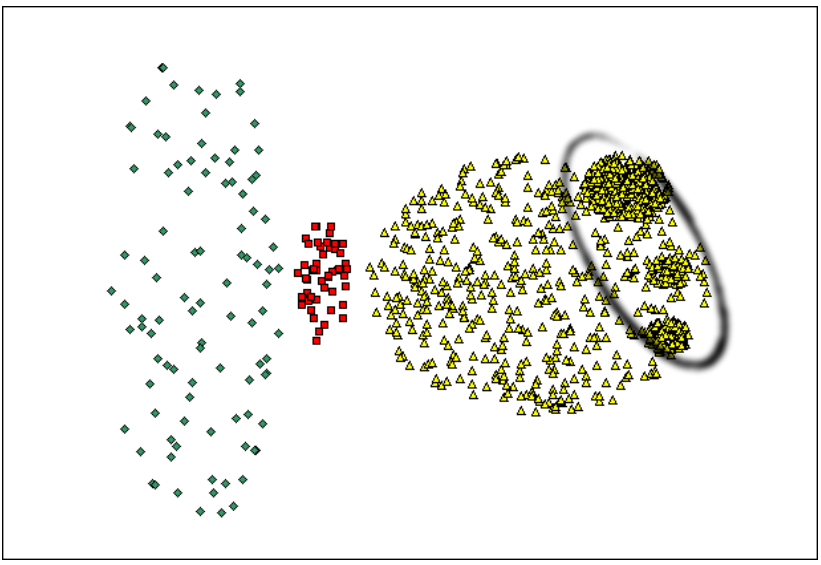
1. Min no of samples required for core point = 48
2. Epsilon (radius) = 83.20858325589253

Now after finding the hyperparameters, we have used sklearn DBSCAN to find the labels of each input data point.

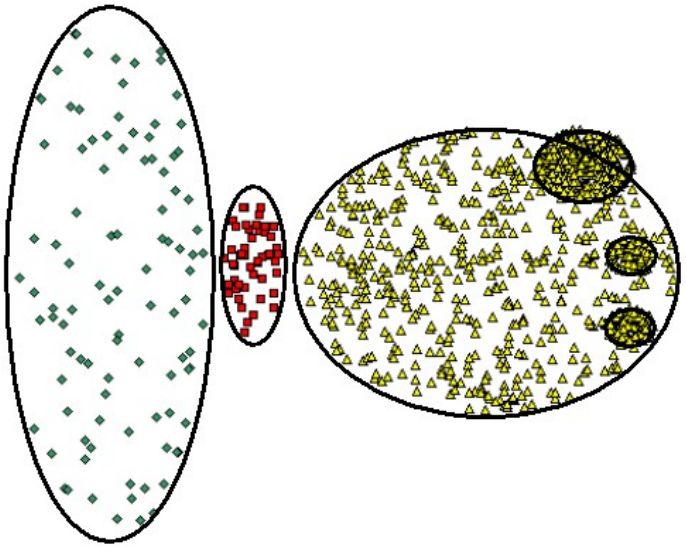
It finds out that there are 20 data points which are abnormal (noise points).

CONCLUSION:

- Finding hyperparameters and proving it's why it works so is hard.
- Data which is more scattered, or data which is interwined with others may not be clustered well.
- Dimension of input data is also a problem.

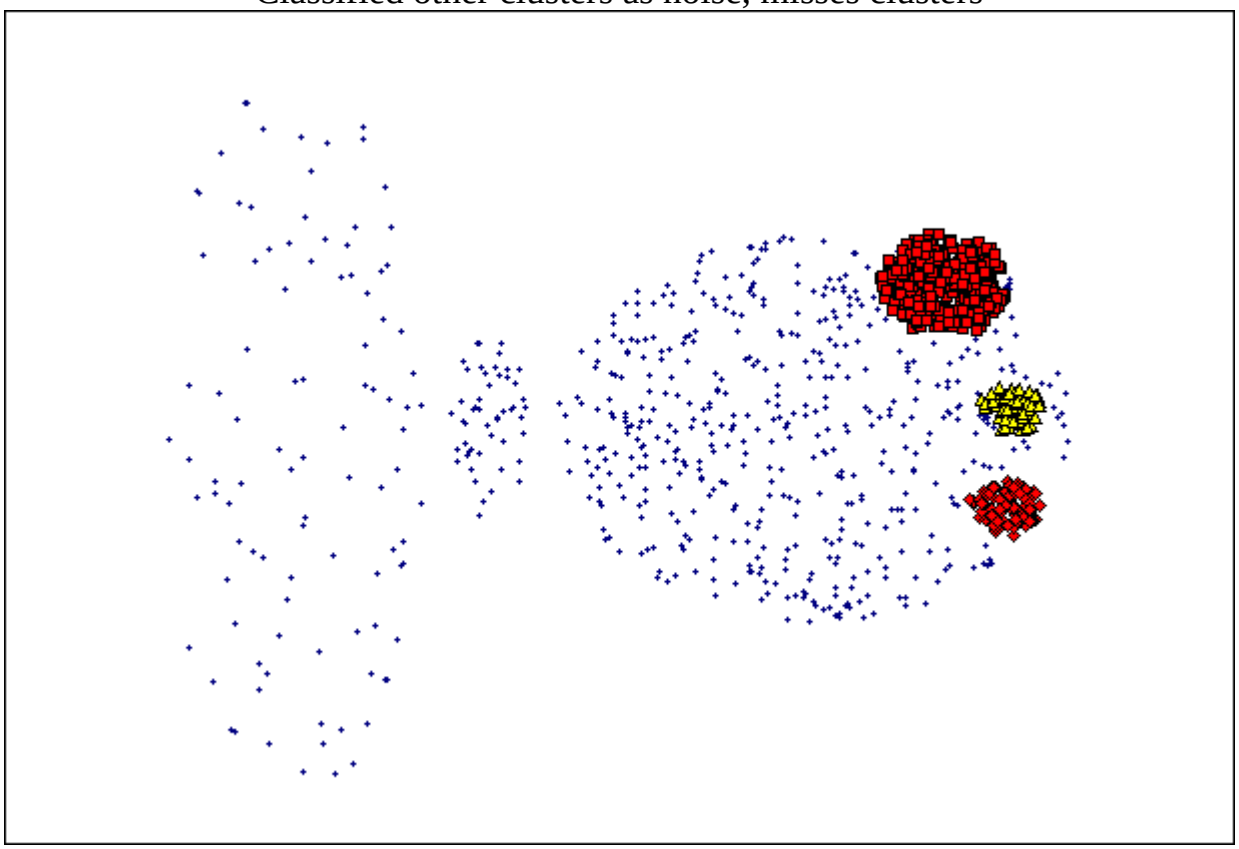


DBSCAN Miss circled clusters



Originial points

Classified other clusters as noise, misses clusters



Source: Sahely mam’s slide